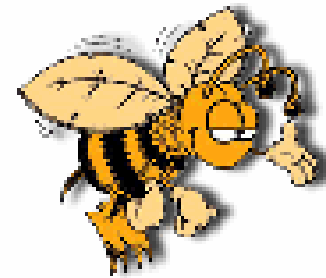




# **MicroArray Database (mAdb) System – Bioinformatics for the Management and Analysis of Spotted and Affymetrix Gene Expression Microarrays**



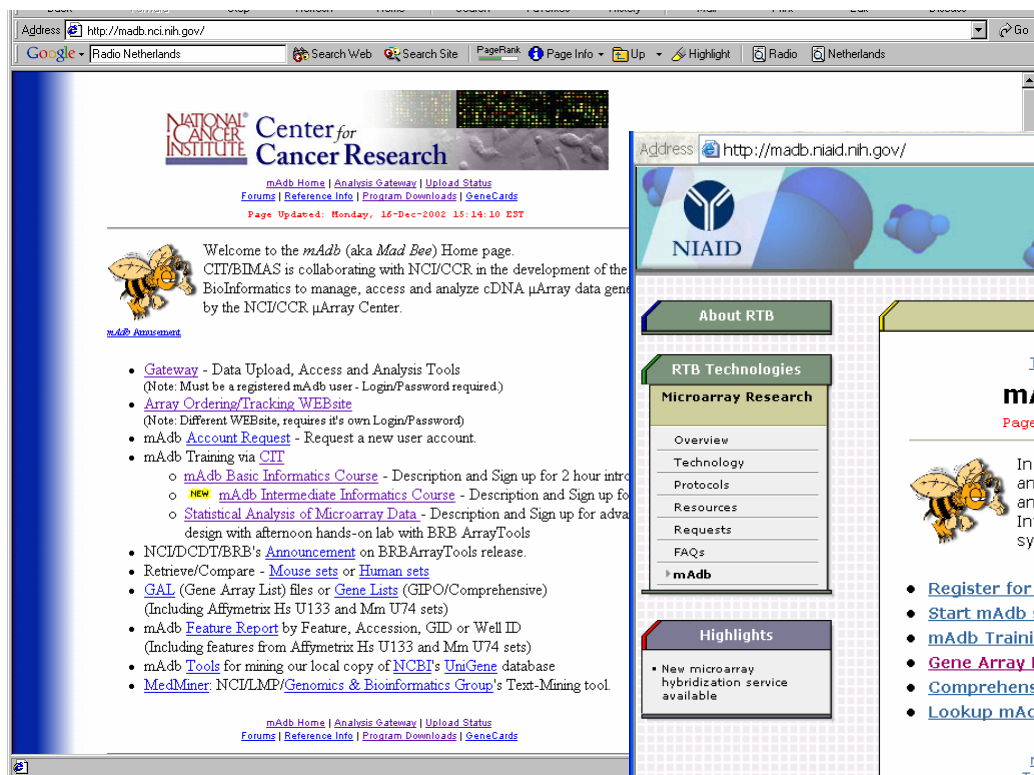
**John M. Greene, Ph.D.**  
*SRA International*  
*ERIC BRC*

**BRC2**  
**May 17, 2005**



# NCI & NIAID mAdb URLs

<http://madb.nci.nih.gov>  
<http://madb.niaid.nih.gov>



Address: <http://madb.nci.nih.gov/>

**NATIONAL CANCER INSTITUTE** Center for Cancer Research

[mAdB Home](#) | [Analysis Gateway](#) | [Upload Status](#)  
[Forums](#) | [Reference Info](#) | [Program Downloads](#) | [GeneCards](#)  
Page Updated: Monday, 16-Dec-2002 15:14:10 EST

Welcome to the *mAdB* (aka *Mad Bee*) Home page.  
CIT/BIMAS is collaborating with NCI/CCR in the development of the Bioinformatics to manage, access and analyze cDNA  $\mu$ Array data generated by the NCI/CCR  $\mu$ Array Center.

[mAdB Announcement](#)

- [Gateway](#) - Data Upload, Access and Analysis Tools  
(Note: Must be a registered mAdB user - Login/Password required)
- [Array Ordering/Tracking WEBSITE](#)  
(Note: Different WEBSITE, requires it's own Login/Password)
- mAdB [Account Request](#) - Request a new user account.
- mAdB Training via [CIT](#)
  - [mAdB Basic Informatics Course](#) - Description and Sign up for 2 hour intro
  - [NEW mAdB Intermediate Informatics Course](#) - Description and Sign up for
  - [Statistical Analysis of Microarray Data](#) - Description and Sign up for advanced design with afternoon hands-on lab with BRB ArrayTools
- NCI/DCIT/BRB's [Announcement](#) on BRBArrayTools release.
- Retrieve/Compare - [Mouse sets](#) or [Human sets](#)
- [GAL](#) (Gene Array List) files or [Gene Lists](#) (GPO/Comprehensive) (Including Affymetrix Hs U133 and Mm U74 sets)
- mAdB [Feature Report](#) by Feature, Accession, GID or Well ID (Including features from Affymetrix Hs U133 and Mm U74 sets)
- mAdB [Tools](#) for mining our local copy of [NCBI's UniGene](#) database
- [MedMiner](#) NCI/LMP/Genomics & Bioinformatics Group's Text-Mining tool.

[mAdB Home](#) | [Analysis Gateway](#) | [Upload Status](#)  
[Forums](#) | [Reference Info](#) | [Program Downloads](#) | [GeneCards](#)

**About RTB**

**RTB Technologies**

**Microarray Research**

Overview  
Technology  
Protocols  
Resources  
Requests  
FAQs  
**mAdB**

**Highlights**

- New microarray hybridization service available



Address: <http://madb.niaid.nih.gov/>

**NIAID** RESEARCH TECHNOLOGIES BRANCH

**Microarray Research**

[mAdB Home Page](#) | [mAdB Gateway](#) | [Upload Status](#)  
[Training/Reference](#) | [Program Downloads](#) | [GeneCards](#)

**mAdB (MicroArray DataBase, a.k.a. "mad bee")**  
Page Updated: Wednesday, 21-Jul-2004 17:08:24 EDT

In collaboration with the Microarray Research Facility at NIAID and the Advanced Technology Center at NCI, the Bioinformatics and Molecular Analysis Section (BIMAS), NIH Center for Information Technology offers the mAdB microarray data analysis system.

- [Register for a mAdB Account](#)
- [Start mAdB session \(requires mAdB account\)](#)
- [mAdB Training/Reference Information](#)
- [Gene Array List \("GAL"\) files for NIAID MRF Arrays](#)
- [Comprehensive Gene Lists](#)
- [Lookup mAdB Features](#)

[mAdB Home Page](#) | [mAdB Gateway](#) | [Upload Status](#)  
[Training/Reference](#) | [Program Downloads](#) | [GeneCards](#)

**NIH Bioinformatics support provided by BIMAS/CBEL/CIT.**  
We can be contacted by [email](#).



# mAdb Background



- Established in 1999 at NIH's CIT to support the NCI microarray printing facilities – SRA awarded contract for staff support
- Goals: Provide an integrated set of web-based analysis tools and a data management system for uploading, analyzing, and maintaining information about the features printed on the chip for cDNA/oligo/Affy Gene Expression data
- Project-oriented design approach to support multiple, independent research projects, using an open systems design, and focusing on 2 color array slides
- System currently supports spotted arrays routinely produced by the two NCI, NIAID, and FDA Microarray Centers - work closely with them
- Currently supports Axon GenePix, Perkin-Elmer QuantArray, Imagene, and Arraysuite II / IP Lab (Yidong Chen, NHGRI) image analysis software for two-color, “Pat Brown-type” spotted arrays
- Affymetrix now available after instruction on needed parameters – limited number of chips supported right now (mouse, human, rat)



## Current CIT mAdb Statistics

- 50,848 Arrays uploaded since Feb. 2000 – now average ~1,100 per month uploaded over last year
- Over one billion cDNA expression measurement points
- ~1,300 registered users (NIH and collaborators worldwide)
- MIAME compliant – MAGE-ML in progress
- Among the largest collections of microarray data in the world, although data sharing is determined by each investigator – no one has access to all the data

# User-configurable mAdb Project Access – Data Security/Sharing

## Add User(s)

mAdb ID# 160 created by "ncidemo" on Jun 26, 2000 at 15:47:00 contains 10 Arrays

**Project Title:** my project

**Description:** Description by jip. Altered @1:00pm on 8/31/2004 and altered again by "easaki" on 9/1/2004

**Comments:** Comments by jip. Altered 8/31/2004

**Access List:** easaki, jmgreene, jpowell, ncidemo

The List below includes **ALL mAdb users** not already having access to this project.

Add User(s)

Reset Form

Cancel

### Check to select User(s) to add to this project

▼ Last name, First name ( Login )

☐ Abdool, Karen ( abdoolk )

☐ Abdullayev, Ziyedulla ( za12 )

☐ Abul-Hassan, Khaled ( hassank )

☐ Ajay, Dr ( ajay\_dr )

☐ Akagi, Keiko ( akagik )

☐ Aksamit, Robert ( aksamit )

☐ Aleman, Claudina ( alemanc )

☐ Alexander, H. Richard ( ralexander )

☐ Ali, Iqbal ( alii )

☐ Alizadeh, Ash ( alizadeh )

☐ Alkharouf, Nawal ( nalkhar )

☐ Amornphimoltham, Panomwat ( pa79w )

▼ Last name, First name ( Login )

☐ Mariotti, jacopo ( mariottj )

☐ Marks-Konczalik, Joanna ( marksj )

☐ Marsh, Katherine ( klmarsh )

☐ Marston, Sarah ( marstons )

☐ Martell, Robin ( rmartell )

☐ Marti, Gerald ( gemarti )

☐ Martin, Kelly ( ktracey )

☐ Martin, Raynaldo ( rmartin )

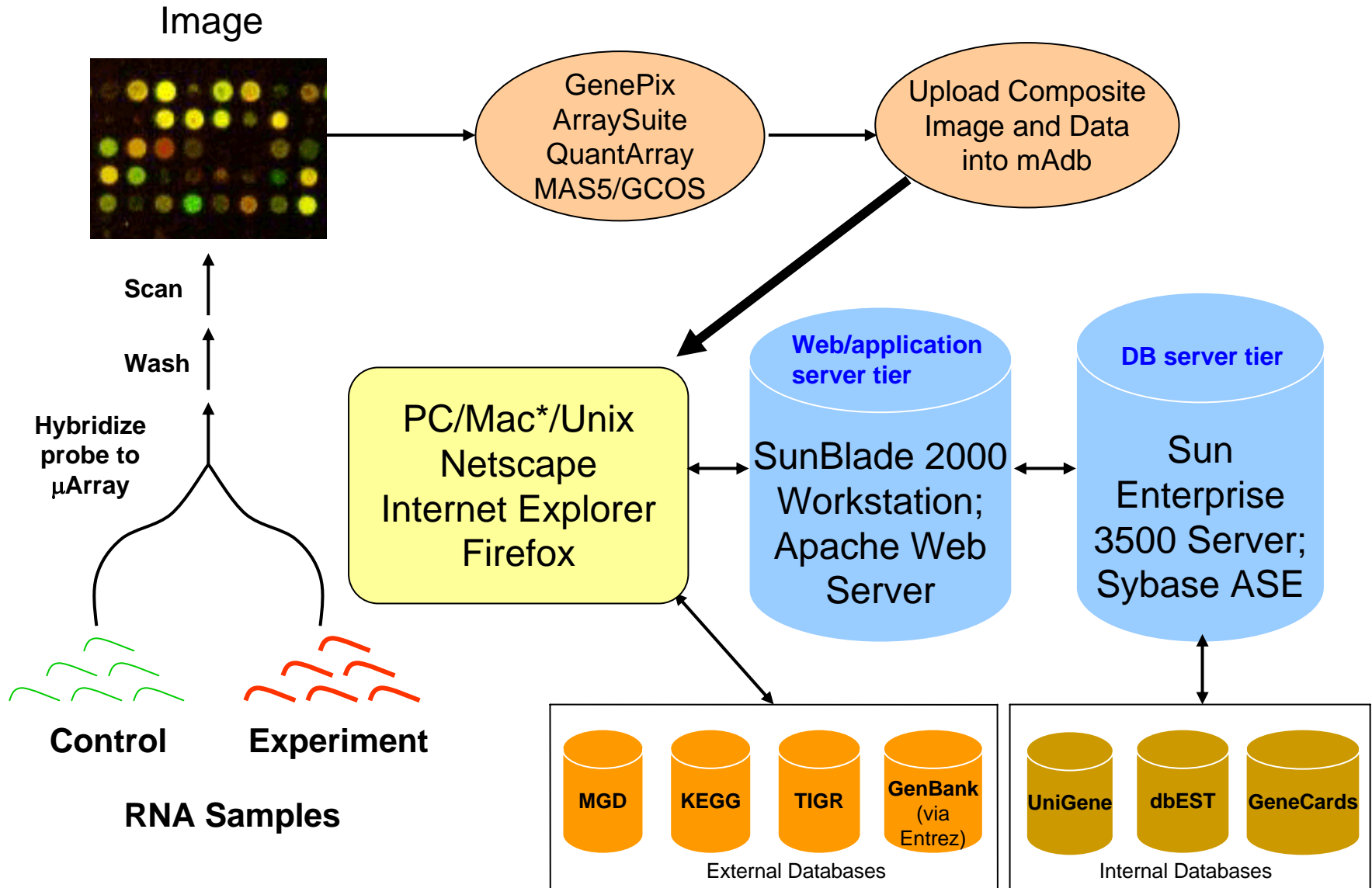
☐ Martinez-Alier, Nuria ( martinezn )

☐ Martinez-Delgado, Beatriz ( martineb )

☐ Mason, Anna ( masona )

☐ Masse, Eric ( massee )

# Physical Architecture for CIT mAdb System







# **mAdb IT Infrastructure at CIT**

## **Web servers:**

- **SunBlade 2000**
- **Open Source Software:**
  - **Apache 1.3 – web server**
  - **Perl 5.8 (Perl/CGI/DBI user interfaces)**
  - **Java applets (visualization tools)**
  - **ImageMagick – image manipulation to allow visualization of individual spots**
  - **R statistical software**

## **Database server:**

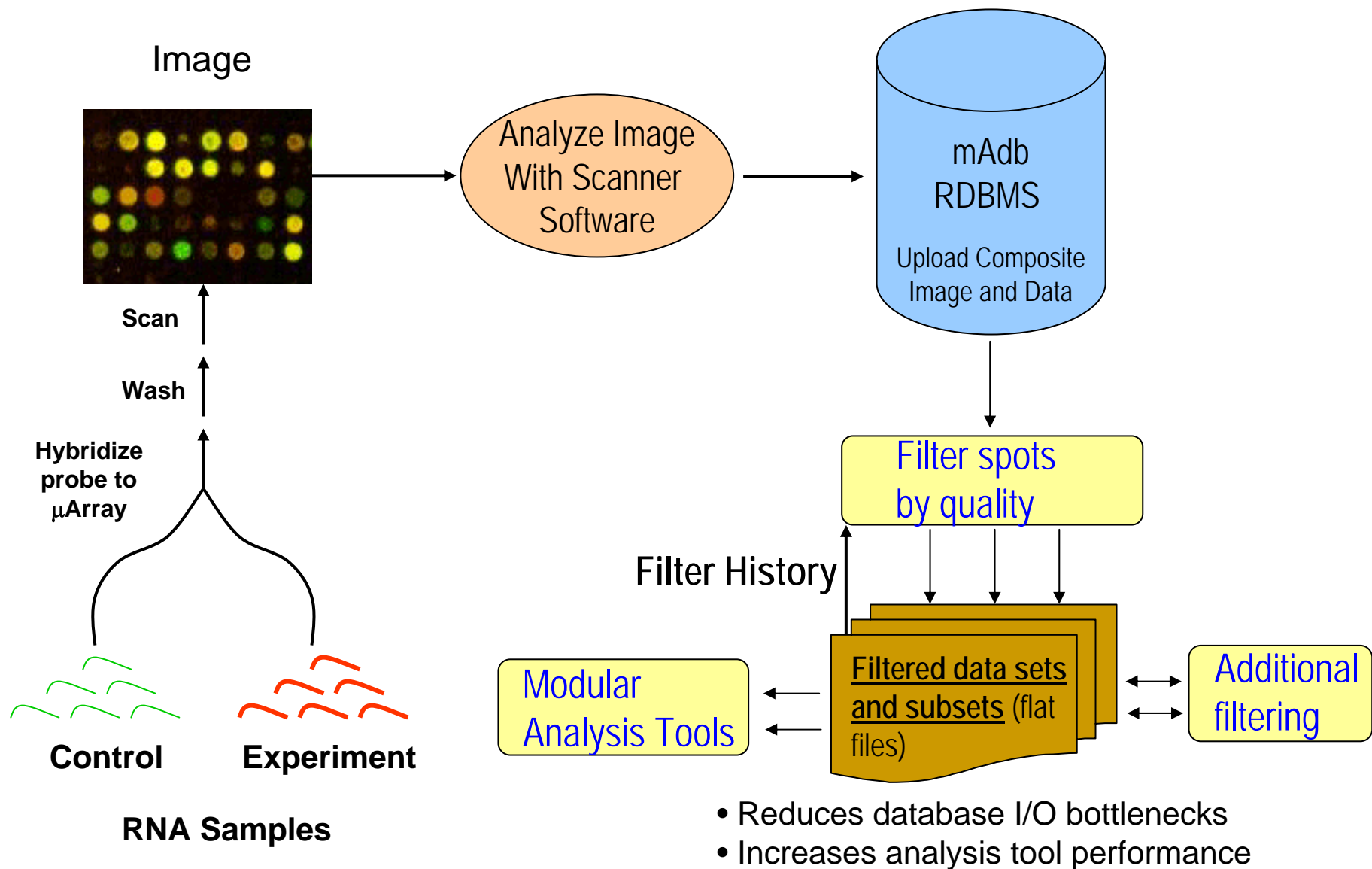
- **Sun Enterprise 3500 with 4 processors**
- **Sybase ASE**

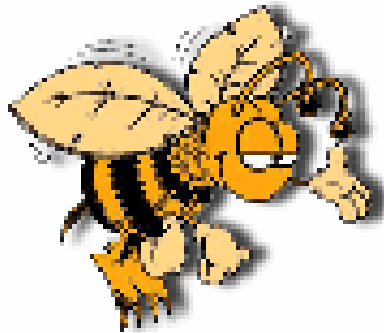
**Storage devices: Sun D2 Storage Arrays 12 x 36 GB**

**Network storage (RAID 5): 1 terabyte**



# Logical Architecture for mAdb System





# Live Demonstration

1. Create a filtered dataset based on array quality parameters
2. Filter out rows missing less than 90% of values to create a subset.
3. Demonstrate a visualization tool – multidimensional scaling



# **mAdb Analysis Paradigm:**

- 1. Create project; Upload arrays to that project**
- 2. Quality control – Project Summary and Graphical Reports**
- 3. Create a filtered dataset:**
  - Extract rows from database
  - Filter spots on quality parameters (spot size, S/N, etc.)
  - Normalize, so different arrays can be compared
  - Align genes from different array layouts (based on well IDs)
- 4. Apply Data/Gene criteria filters, if desired, to create subset dataset(s)**
- 5. Apply appropriate Analysis/Visualization Tools to the dataset(s)**
- 6. Repeat Steps 3, 4, and 5 as desired**
- 7. Interpret Datasets/Results**

# Uploading Arrays via the Web

## Add a New Array Experiment for Project: MYTEST PROJECT

### Experiment Information

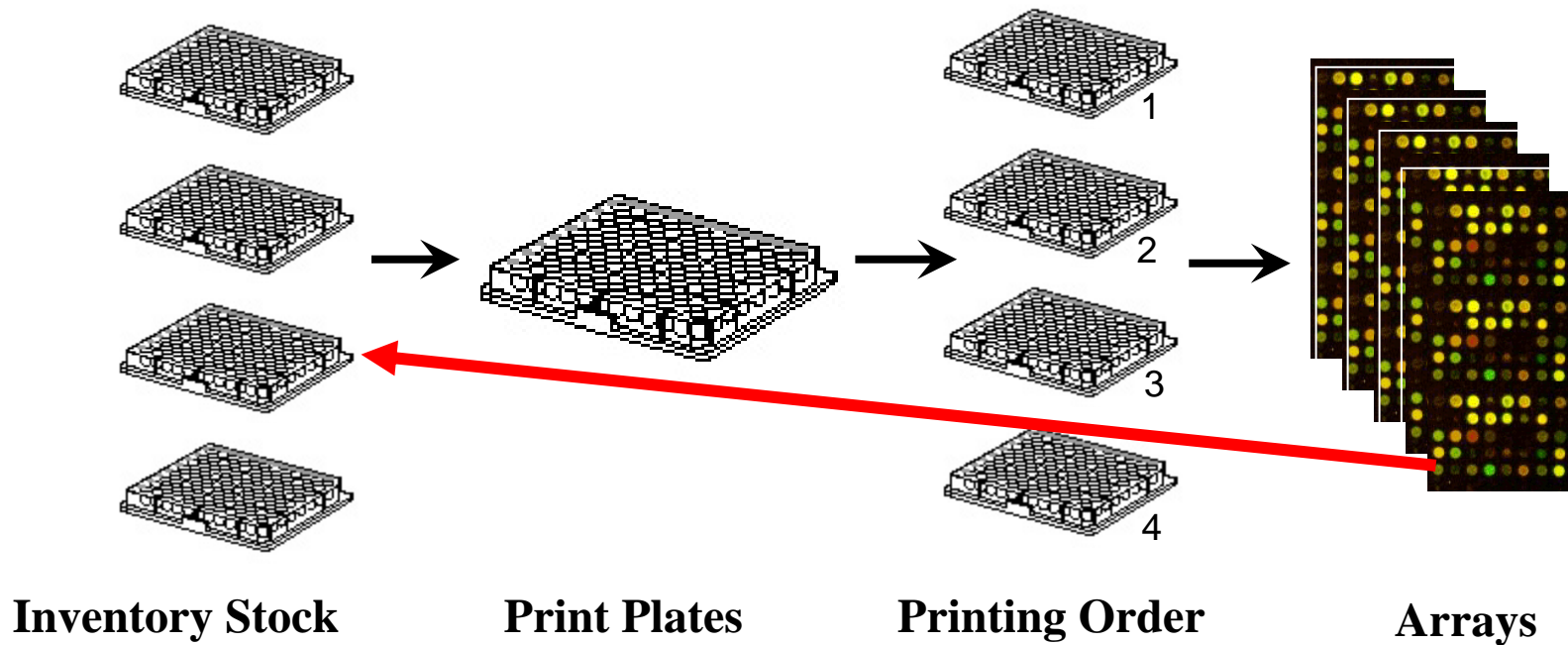
Array Print Set:	<input type="text" value="Hs-OC-2p13-120699"/>	
Array Name:	<input type="text" value="HsOC2p13-46"/>	Suggested form: HsOC2p13-45
Short Description:	<input type="text" value="Colchicine time course - 1 hr"/>	
Long Description:	<input type="text"/>	
Probe:	Channel A (generally <b>Cy3</b> tagged) <input type="text" value="Untreated control"/>	Channel B (generally <b>Cy5</b> tagged) <input type="text" value="1 hr experimental"/>
Probe Label:	<input type="text" value="Cy3"/>	<input type="text" value="Cy5"/>

### Composite Image & Arraysuite Sample Intensities or GenePix GPR Files

Image File:	<input type="text" value="myarray.jpg"/>	<input type="button" value="Browse..."/>
Data File:	<input type="text" value="myarray.gpr"/>	<input type="button" value="Browse..."/>

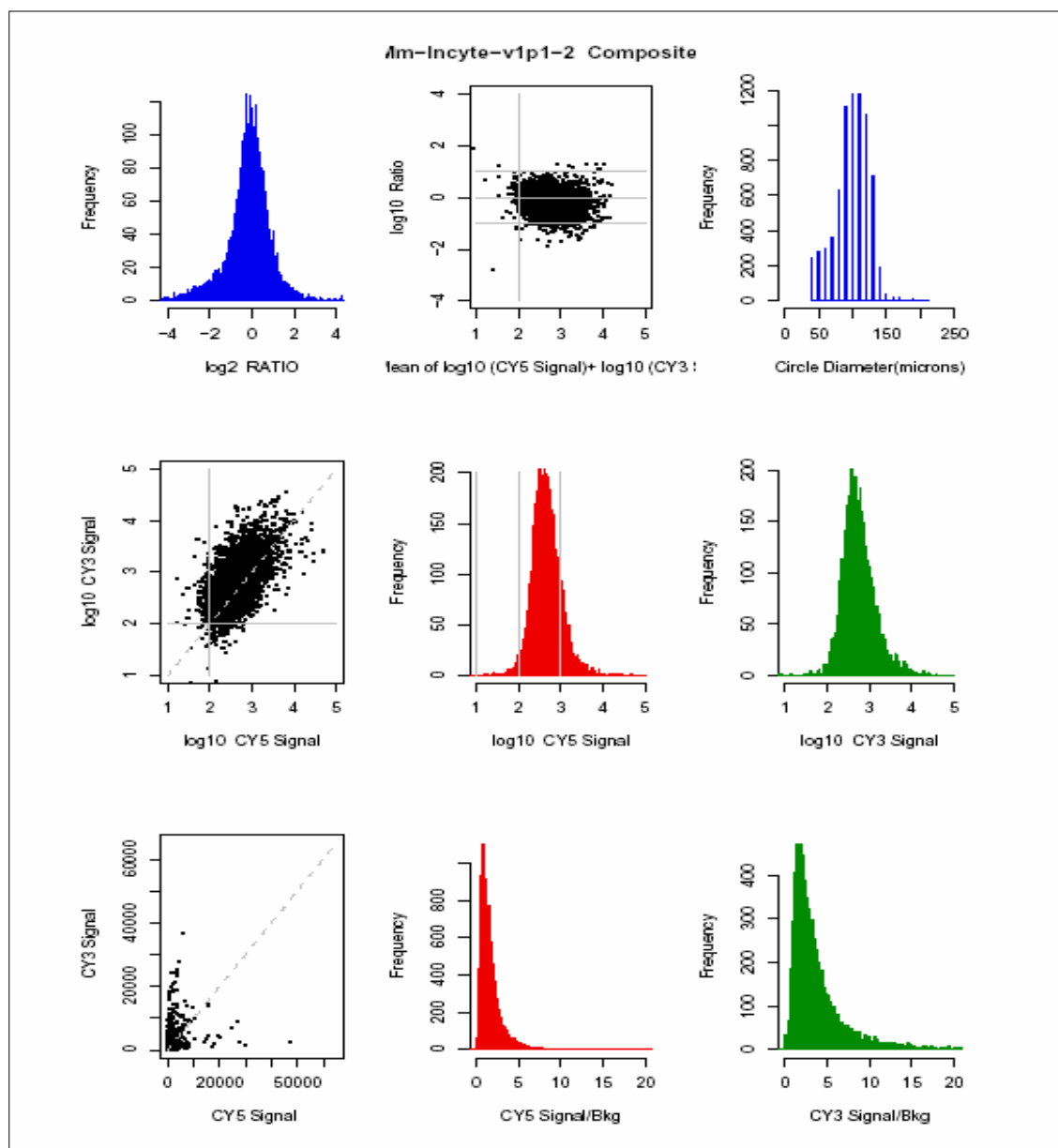
[Return to Data Loading Page](#)

# Feature Tracking



Array spots are traced back to inventory stock wells (well ID). This well ID is one of the primary methods used to align results across array sets, and allows mAdb staff to correct printing errors.

# Array Quality Report Graphics



# Data Extraction Tool – creates filtered datasets

Applies spot filtering options to selected arrays and creates a new working dataset.

**Signal, Normalization & Ratio Options**

Signal Calculation: **Mean Int - Median Bkg**

Normalization Method: **50th Percentile (Median)**

Default Ratio: **ChanB/ChanA (Cy5/Cy3)**

☐ Limit Normalization to HouseKeeping Genes

**Caution:** Most array prints do not have an identified set of HouseKeeping Genes

**Spot Filter Options**

Check boxes on the left to activate specific criteria

☒ Exclude any Spots Flagged as **Bad or NF**

☐ Target diameter is between **70**  $\mu\text{m}$  and **180**  $\mu\text{m}$

	Chan A (cy3)		Chan B (cy5)
<input type="checkbox"/> Target Pixels 1 SD above Bkg $\geq$	<b>60</b> %	and	<b>60</b> %
<input type="checkbox"/> Signal/Background Ratio $\geq$	<b>2</b>	and	<b>2</b>
<input checked="" type="checkbox"/> Signal $\geq$	<b>50</b>	and	<b>50</b>
<input type="checkbox"/> Override if Chan B Signal $\geq$			<b>2500</b>
<input type="checkbox"/> Override if Chan A Signal $\geq$	<b>5000</b>		

**Dataset Properties**


Rows Ordered by: **Descending**

Dataset Location: **Temporary Area**

Dataset Label: (Optional)



# Reconfigurable Data Display Output

Data Retrieval & Display Options 

Retrieve Data Set formatted for Eisen Cluster Program































Retrieve Data set formatted for MS-Excel
  
☐ Apply log2 transform

Redisplay
  
☐ Show Array Details at the top of the page
   
 Background Color Red/Yellow/Green Contrast 1
  
 Limiting display to to 25 genes
  

☒ Show Data Values
   
☒ Apply log2 transform
   
☒ Show Spot Images
   
☐ Show Map Information
   
☐ Show BioCarta Pathways
   
☐ Show GO Tier 2 Component
   
☐ Show GO Tier 2 Function
   
☐ Show GO Tier 2 Process
   
☒ Show Gene Description
   
☒ Show Average(Log2 Ratio)
   
☐ Show Variance

☐ Use Names in Column Heading
   
☐ Use Description in Column Heading
   
☐ Show Gene Symbols
   
☐ Show UniGene Cluster
   
☐ Show KEGG Pathways
   
☐ Show GO Tier 3 Component
   
☐ Show GO Tier 3 Function
   
☐ Show GO Tier 3 Process
   
☐ Show GO Terms
   
☐ Show Max(Log2 Ratio)-Min(Log2 Ratio)

➔ Records 1 to 25 of 741 total records displayed.

A	A	A	A	A	A	A	A	A	A				
#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Aver	Well ID	Feature ID	Description
										-0.7587	613012	<a href="#">IMAGE:791264</a>	ATX1 (antioxidant protein 1) homoc
-0.4974	-0.2365	-1.2837	-0.3389	-0.6204	-0.6047	-1.5558	-1.3646	-0.6615	-0.4235				
										-0.3085	613034	<a href="#">IMAGE:761319</a>	Mus musculus, clone MGC:36382
0.4939	-0.4061	-0.8428	-0.3872	-0.3727	-0.5507	-0.4590	0.2002	-0.2750	-0.4855				
										0.9069	613040	<a href="#">IMAGE:933491</a>	RIKEN cDNA 2310044F10 gene
0.9162	0.9714	0.6831	0.8959	0.8962	1.1806	0.1383	0.9012	1.2643	1.2219				

# Feature Report - Integration of Gene Information

**Feature Report - Microsoft Internet Explorer**

File Edit View Favorites Tools Help Links »

## *mAdb* Feature Report

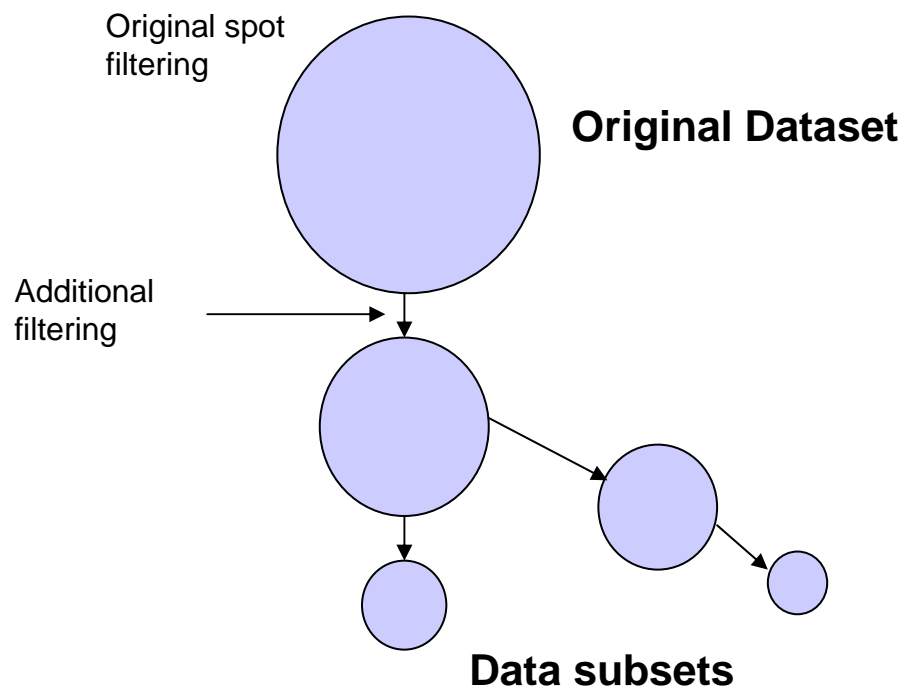
---

<b>Clone</b>	<a href="#">IMAGE:301551</a>																																
<b>Library Source</b>	Soares_fetal_lung_NbHL19W																																
<b>Sequence Verification</b>	Unknown																																
<b>Annotated Simple PID</b>	Integrin, alpha V (vitronectin receptor, alpha polypeptide, antigen CD51)																																
<b>Annotated NG Assignment</b>	<a href="#">M14648</a> Human cell adhesion protein (vitronectin) receptor alpha subunit mRNA, complete cds																																
<b>Annotated Categories</b>	Adhesion																																
<b>5' Sequence</b>	<a href="#">W17002</a> UCSC's <a href="#">GenomeViewer</a>																																
<b>5' UG Title</b>	integrin, alpha V (vitronectin receptor, alpha polypeptide, antigen CD51)																																
<b>5' UG Cluster</b>	<a href="#">TP Hs.295726</a> NCBI's <a href="#">LocusLink</a> Stanford's <a href="#">S.O.U.R.C.E.</a>																																
<b>5' UG Gene</b>	ITGAV <a href="#">GeneCards</a> <a href="#">MedMiner</a> NCBI's <a href="#">Map Viewer</a>																																
<b>5' UG LL Summary</b>	ITAGV encodes integrin alpha chain V. Integrins are heterodimeric integral membrane proteins composed of an alpha chain and a beta chain. The I-domain containing integrin alpha V undergoes post-translational cleavage to yield disulfide-linked heavy and light chains, that combine with multiple integrin beta chains to form different integrins. Among the known associating beta chains (beta chains 1,3,5,6, and 8; 'ITGB1', 'ITGB3', 'ITGB5', 'ITGB6', and 'ITGB8'), each can interact with extracellular matrix ligands; the alpha V beta 3 integrin, perhaps the most studied of these, is referred to as the Vitronectin receptor (VNR). In addition to adhesion, many integrins are known to facilitate signal transduction.																																
<b>5' UG Ontology</b>	<table border="0"> <tr> <td><a href="#">GO</a> <sup>™</sup> Annotations</td> <td><a href="#">Evidence</a> </td> <td>Source</td> <td>Pub</td> </tr> <tr> <td>• <a href="#">cell adhesion</a></td> <td>P</td> <td>Proteome</td> <td><a href="#">PM</a></td> </tr> <tr> <td>• <a href="#">cell adhesion receptor</a></td> <td>P</td> <td>Proteome</td> <td><a href="#">PM</a></td> </tr> <tr> <td>• <a href="#">integral plasma membrane protein</a></td> <td>P</td> <td>Proteome</td> <td><a href="#">PM</a></td> </tr> <tr> <td colspan="4">Other Annotations</td> </tr> <tr> <td>• Integral membrane</td> <td>NR</td> <td>Proteome</td> <td><a href="#">PM</a></td> </tr> <tr> <td>• Receptor (signalling)</td> <td>NR</td> <td>Proteome</td> <td><a href="#">PM</a></td> </tr> <tr> <td>• Control of Cell Proliferation</td> <td>E</td> <td>Proteome</td> <td><a href="#">PM</a></td> </tr> </table>	<a href="#">GO</a> <sup>™</sup> Annotations	<a href="#">Evidence</a>	Source	Pub	• <a href="#">cell adhesion</a>	P	Proteome	<a href="#">PM</a>	• <a href="#">cell adhesion receptor</a>	P	Proteome	<a href="#">PM</a>	• <a href="#">integral plasma membrane protein</a>	P	Proteome	<a href="#">PM</a>	Other Annotations				• Integral membrane	NR	Proteome	<a href="#">PM</a>	• Receptor (signalling)	NR	Proteome	<a href="#">PM</a>	• Control of Cell Proliferation	E	Proteome	<a href="#">PM</a>
<a href="#">GO</a> <sup>™</sup> Annotations	<a href="#">Evidence</a>	Source	Pub																														
• <a href="#">cell adhesion</a>	P	Proteome	<a href="#">PM</a>																														
• <a href="#">cell adhesion receptor</a>	P	Proteome	<a href="#">PM</a>																														
• <a href="#">integral plasma membrane protein</a>	P	Proteome	<a href="#">PM</a>																														
Other Annotations																																	
• Integral membrane	NR	Proteome	<a href="#">PM</a>																														
• Receptor (signalling)	NR	Proteome	<a href="#">PM</a>																														
• Control of Cell Proliferation	E	Proteome	<a href="#">PM</a>																														
<b>5' UG RefSeq</b>	<a href="#">NM_002210</a>																																
<b>5' UG Cytoband</b>	2q31-q32																																
<b>5' Submitted PID</b>	gb:M14648 VITRONECTIN RECEPTOR ALPHA SUBUNIT PRECURSOR (HUMAN);																																

Internet

# Additional Data Filtering Options

Applies selected filtering options to the dataset based on values in the data and creates a new subset.



Check boxes on the left to activate specific filters ▼

**Missing Value Filters** 🎧

☒ Genes: Require values in  $\geq$   % Arrays

☐ Arrays: Require values in  $\geq$   % Genes

**Gene Filters** 🎧

☐ Ratio  $\geq$   in  $\geq$   Arrays  ☒ Apply Symmetrically

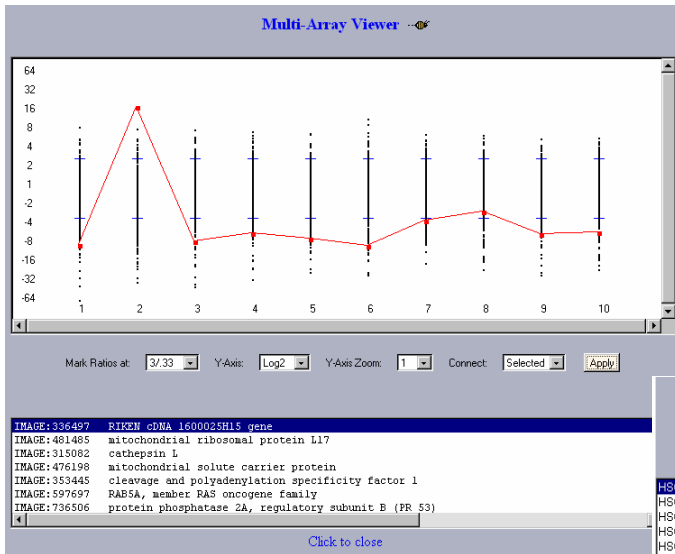
☐ Ratio  $\geq$   in  $\geq$   Arrays  OR  
Ratio  $\leq$   in  $\geq$   Arrays

☐ Average Ratio  $\geq$   ☐ Apply Symmetrically

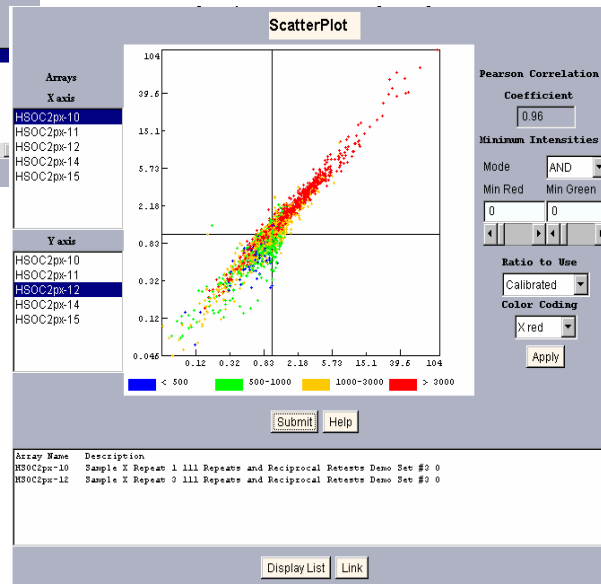
☐ Max (Ratio) / Min (Ratio)  $\geq$

☒ Variance (Gene Vector) percentile  $\geq$   %

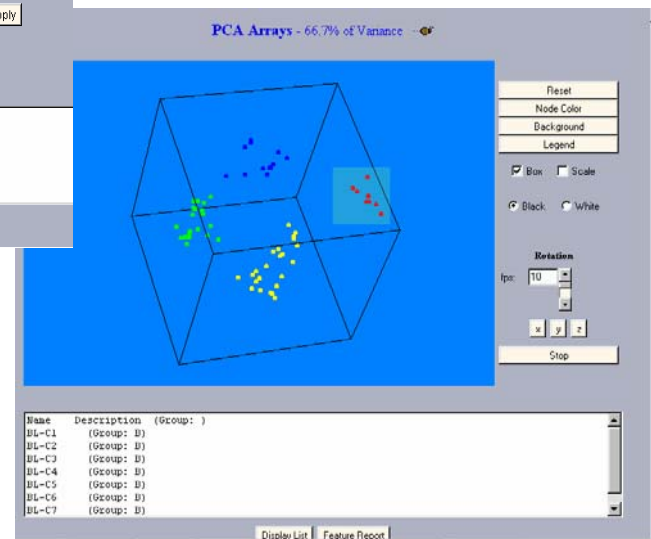
# Visualization Tools for Microarray Datasets



Multiple array graphical viewer



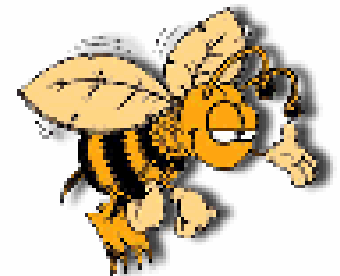
Interactive scatter plot



PCA and MDS – 3D viewer

## **Features and analysis tools currently included in the mAdb system are:**

- **Agglomerative hierarchical, K-means, and self-organizing map (SOM) clustering**
- **Principal Components Analysis (PCA) and Multidimensional Scaling (MDS)**
- **Scatter plot and Multiple array graphical viewers**
- **PAM (Prediction Analysis of Microarrays) classifier from Stanford**
- **Boolean comparison of datasets**
- **Array group assignment and averaging**
- **t-tests, Wilcoxon Rank Sum, ANOVA, and Kruskal-Wallis statistical analyses**
- **Pathways Summary reports (BioCarta, KEGG, Stanford GO)**
- **Correlation Summary report, Array Quality Graphic Report**
- **Configurable data display**
- **Ability to refresh gene information**
- **History of filtering of data subsets**
- **Export of data to Excel, tab-delimited files, GeneSpring format**
- **Keyword query**



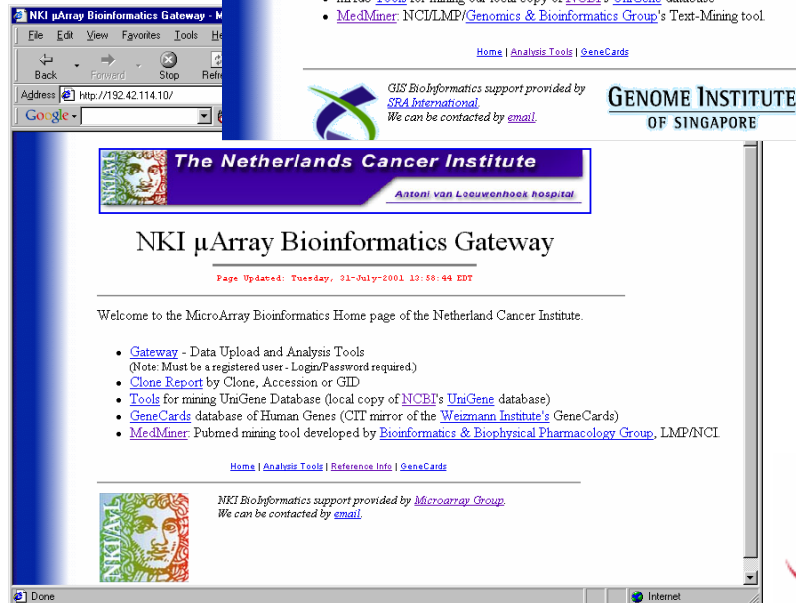
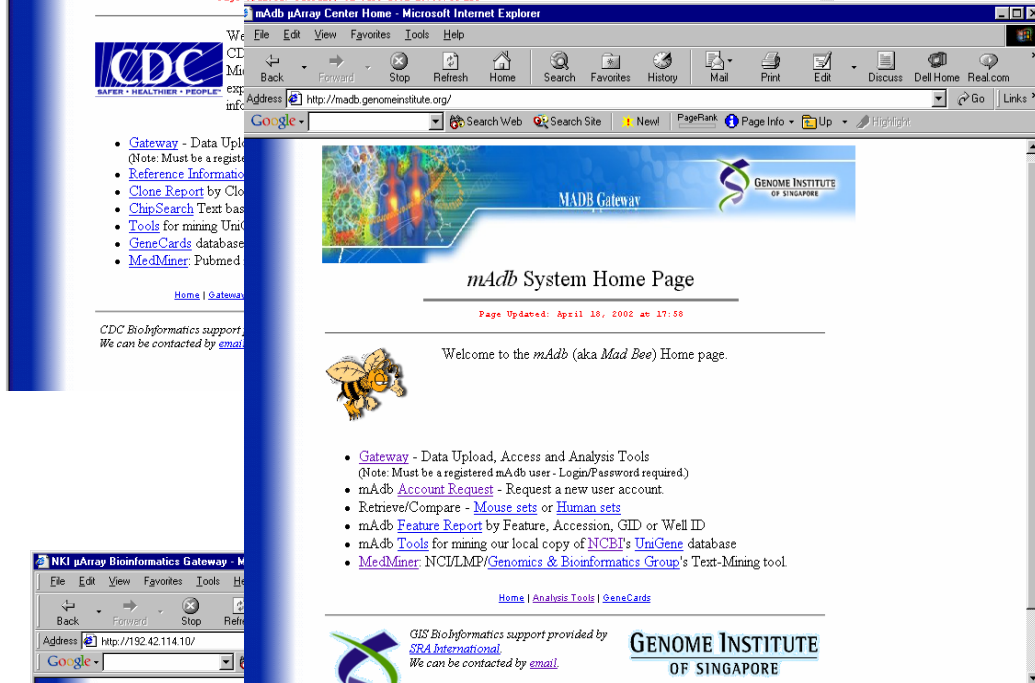
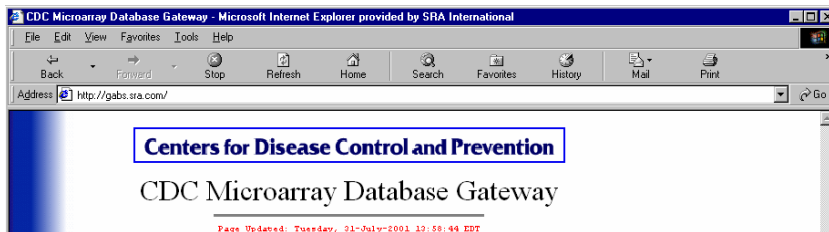
## **5/6/2005 - What's New on mAdb**

- **"Group Comparison (t-test, ANOVA, Wilcoxon,...)"** replaces **"Two or More Group Comparison"** tool, includes previous functionality, plus support for single group tests.
- **"Group Statistics (mean, median, stddev...)"** - calculates statistics values for each Group.
- **"SAM: Significance Analysis of Microarrays"** - support for SAM analysis.

# Issues for BRC use

- **Gov't owned – ERIC cannot Open Source it per se**
- **Cf. Owen yesterday – mAdb needs testing, packaging, and documentation**
- **To date, mAdb has not dealt with multiple organisms' DNA on same chip**
- **More uploading automation – need to add chip layout input from users**
- **Can make accounts available once up on ERIC for BRCs to try out....**





To date, the NCI/CIT mAdB system has been replicated by SRA International for the:

- Genome Institute of Singapore
- CDC, with concurrent development of an epidemiological database and additional statistical tool refinements.
- Netherlands Cancer Institute, with expansion of import capability
- NINDS, with conversion to Oracle

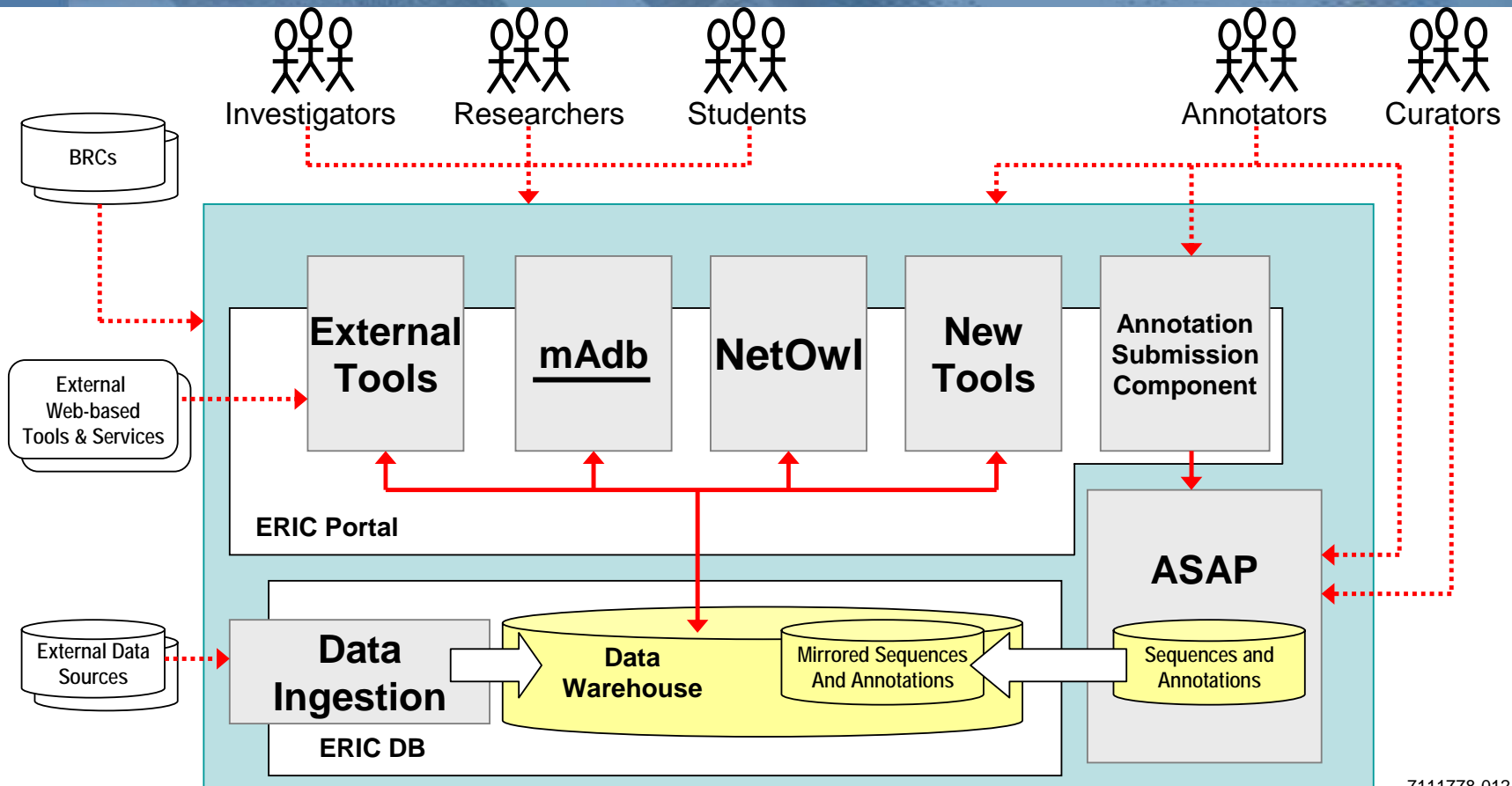
The NCI system is also being used by investigators at NIAID, FDA/CBER, NIMH, and other intramural NIH programs.





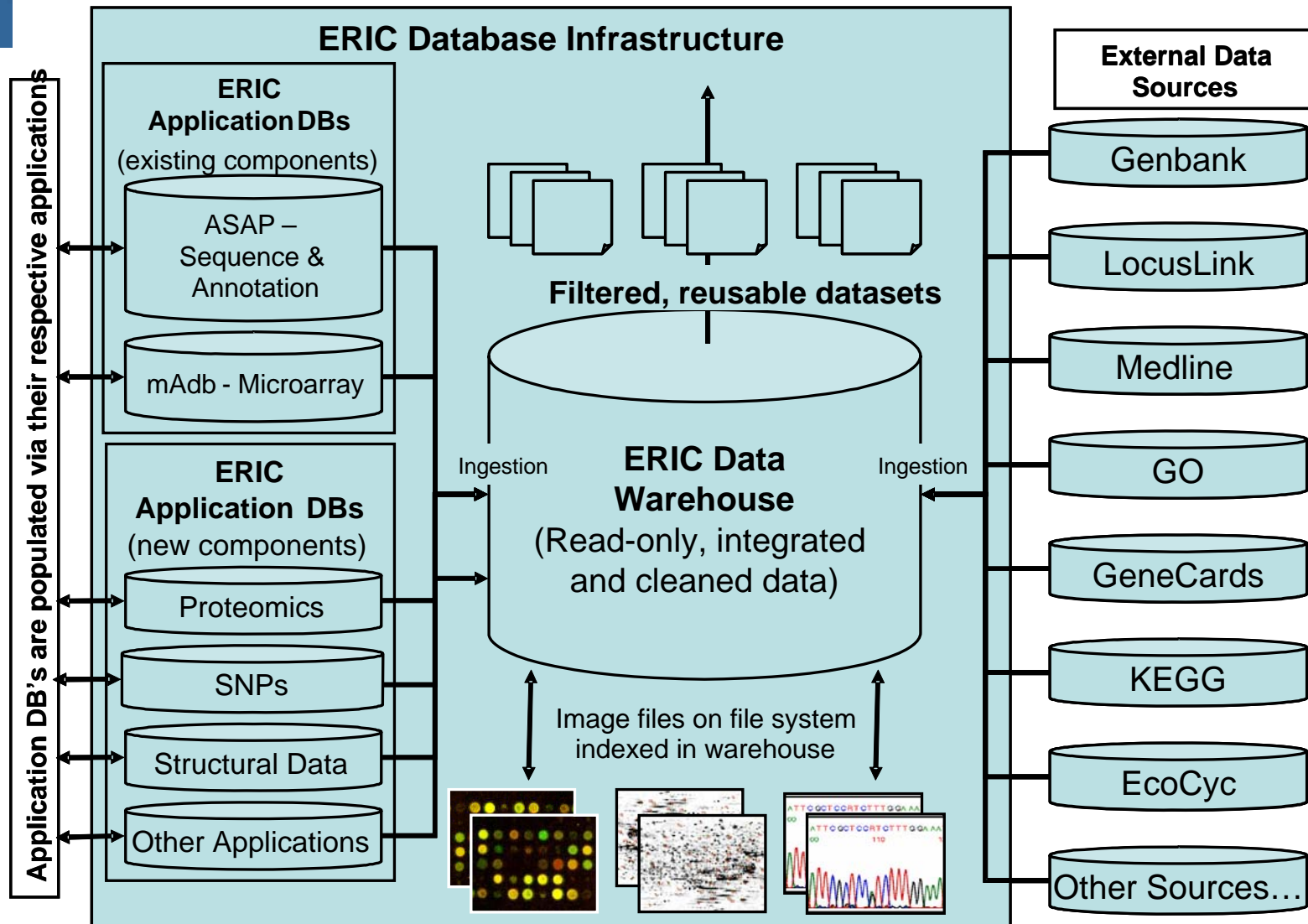
# Enteropathogen Resource Integration Center

## Bioinformatics Resource Center



7111778-012.ppt

ERIC is a **pathogen-centric**, portal based system. mAdb is one component of the system, and work on replication of the mAdb system for ERIC is beginning now.



ERIC will make use of a **data warehouse** approach to store both contributed pathogen data (annotations, sequence, microarray, proteomics, etc.) and data from external sources. This will enable better integration of inputs from many sources.

## **CIT mAdb Development and Support Team:**

- **John Powell, Chief, BIMAS, CIT**
- **Liming Yang, Ph.D.**
- **Jim Tomlin**
- **Lynn Young, Ph.D.**
- **Esther Asaki\***
- **Yiwen He, Ph.D.\***
- **Kathleen Meyer\***
- **Tim Ruppert\***

**\*SRA International contractor**

- **John Powell received an NIH Director's Award for Development and Support of the mAdb System**
- **2004 SRA Project Team Excellence Award – NCI mAdb Project Team**



# ERIC Team:



## Scientific Co-Directors:



John Greene, Ph.D. - PI and Project Director



Nicole Perna, Ph.D. - Scientific Co-Director

Fred Blattner, Ph.D. - Scientific Co-Director

## ERIC Curators:



Guy Plunkett III, Ph.D. - Senior Curator; David Bowen, Ph.D.;

Val Burland, Ph.D.; Eric Cabot, Ph.D.; Jeremy Glasner, Ph.D.

## ERIC Technical Team:



Matt Shaker; **Robin Martell**; **Tom Hampton**; Lorie Shaul; Panna Shetty;

**Mary Wong**, Mark Backus



Paul Liss; Michael Rusch



<http://www.ericbrc.org>

[info@ericbrc.org](mailto:info@ericbrc.org)